Is big data ready TODAY to improve patient outcomes or is it a new generation of garbage in/garbage out?

Gordon B. Mills
University of Texas, MD Anderson Cancer Center, USA
POTENTIAL CONFLICT OF INTEREST DISCLOSURES

• Financial Relationships
  – **SAB/Consultant:** Adventist Health, Allostery, AstraZeneca, Catena Pharmaceuticals, Critical Outcome Technologies, ImmunoMET, Lilly, Medimmune, Nuevolution, Novartis, Precision Medicine, Provista Diagnostics, Roche, Signalchem Lifesciences, Symphogen, Takeda/Millenium Pharmaceuticals, Tau Therapeutics, Tarveda,
  – **Stock/ Options/Financial:** Catena Pharmaceuticals, ImmunoMet, Spindle Top Ventures
  – **Licensed Technology** HRD assay to Myriad Genetics
  – **Sponsored Research:** Adelson Medical Research Foundation, AstraZeneca, Breast Cancer Research Foundation, Critical Outcomes Technology, Illumina, Ionis, Karus, Komen Research Foundation, Nanostring, Takeda/Millenium Pharmaceuticals

I will not discuss off label use and/or investigational use of drugs
BIG DATA

- The present
- The future
- Monetization
- Business Plan
- Big Business
- Buzz Word
- Hype
- Fund raising

From Turing (104th birthday) to the Big data Deluge
Big Data
Garbage in Garbage out

• Not a new concept
  – Coined in November 10, 1957, syndicated newspaper article about US Army mathematicians and their work with early computers,[1] in which an Army Specialist named William D. Mellin explained that computers cannot think for themselves, and that "sloppily programmed" inputs inevitably lead to incorrect outputs.

• What is new
  – Big data is messy data
  – More messy data: We can really accumulate lots of garbage
  – If there is enough messy data and the right computer program (magical black box) “something” useful will fall out
Not all garbage is useless
Garbage can be useful

Tourist industry in Churchill, Manitoba
Why data bases should not replace randomized clinical trials.  
Byar DP.  
Advances in computer technology have made it possible to store large amounts of observational data concerning treatment of patients for medical disorders. It has been suggested that these data banks might replace randomized clinical trials as a means of evaluating the efficacy of therapies. A review of the methodological problems likely to arise in analyzing such data for the purpose of comparing treatments suggests that sound inferences would not generally be possible because of difficulties with bias in treatment assignment, nonstandard definitions, definitions changing in time, specification of groups to be compared, missing data, and multiple comparisons.
Big Data Mining → Big Data Mining Accidents
David P. Byar 1938-1991

BIG DATA TOLD ME TO DO IT

UNLESS YOU PAY ATTENTION TO HISTORY YOU ARE DOOMED TO REPEAT IT
BIG DATA

• Concede to my esteemed colleague that big data has incredible potential to contribute to improved patient outcomes in the future if we do it right.

Calvin Coolidge

The most common commodity in this country is unrealized potential
UNREALIZED POTENTIAL IS PAINFUL
Erwin Gelfand PDF Supervisor
Integrative Analysis
The TCGA Research Network

- Tumor genotyping (mutations, sequencing, amplifications, deletions)
- Tumor phenotype
- Tumor cells
- Stroma cells
- Lymphocytes
- Expression profiling (mRNA, miRNA)
- Epigenetic (Methylation of DNA)
- Clinical Outcomes

Functional Proteomics
The Cancer Genome Atlas
Single biggest problem with the TCGA is the failure to have equivalent high quality patient data
Many “marker” papers do not include correlations with outcomes
YOU DON’T WANT TO LOOK BACK AND KNOW YOU COULD’VE DONE BETTER
Big Data is Messy Data

• Questions must be robust to messy data
  – Should we change patient management based on messy data

• Are patient and omic data robust to messiness

The challenges of big data

Elaine R. Mardis*

This editorial outlines several such challenges as a means of illustrating that the path to big data revelations is paved with perils that the scientific community must overcome to pursue this important quest
Genome Data Commons

- The NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine.

- As a growing knowledge system for cancer, the GDC also enables researchers to submit data, and harmonizes these data for import into the GDC. As more researchers add clinical and genomic data to the GDC, it will become an even more powerful tool for making discoveries about the molecular basis of cancer that may lead to better care for patients.

- NO FUNDING FOR OBTAINING AND ENTERING HIGH QUALITY CLINICAL DATA IN A USABLE FORMAT
Big Data is Messy Data

• Is clinical data including outcomes robust to messiness

• Molecular epidemiology
  – Not robust to messiness

• Clinical trials
  – Not robust to messiness
  – Many phase II trials fail to predict phase III trials

• Need to invest in obtaining, curating, and sharing high quality molecular data linked to even higher quality patient data

• Who will pay for it: Will academics have functional access
Big data is big business

THE PROJECTED RISE OF BIG DATA APPS AND ANALYTICS

As the world becomes more and more digitized, the amount of big data out there will only increase in size over time. As it is, we’ve seen steady growth in all areas—but the Apps and Analytics of big data may see the sharpest spike in the years to come.

According to Statista, the market volume of big data in Apps and Analytics is projected to increase from 2.84 billion to 8.33 billion by 2020. Knowing this spike is coming, it’s imperative for businesses to get a handle on upcoming big data trends now.

Monetization
Business Plan
Sustainability
Who will pay for it

Democratization
Steve Friend
Community Access
Big data is big business
The Jeff Drazen Controversy

DATA SHARING
The aerial view of the concept of data sharing is beautiful. What could be better than having high-quality information carefully reexamined for the possibility that new nuggets of useful data are lying there, previously unseen? The potential for leveraging existing results for even more benefit pays appropriate increased tribute to the patients who put themselves at risk to generate the data. The moral imperative to honor their collective sacrifice is the trump card that takes this trick.
The Jeff Drazen Controversy

• A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group’s data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as “research parasites.”
SYSTEMS (SYSTEMATIC) APPROACH TO PERSONALIZED CANCER THERAPY

Gordon B Mills
MD Anderson Cancer Center
SYSTEMS BIOLOGY IS A DATA DRIVEN SCIENCE
Lee Hood
SYSTEMS (SYSTEMATIC) APPROACH TO PERSONALIZED CANCER THERAPY

SYSTEMS BIOLOGY IS A DATA DRIVEN SCIENCE
Lee Hood

DATA PARASITES
Jeff Drazen NEJM
SYSTEMS (SYSTEMATIC) APPROACH TO PERSONALIZED CANCER THERAPY

SYSTEMS BIOLOGY IS A DATA DRIVEN SCIENCE
Lee Hood

DATA PARASITES
Jeff Drazen NEJM

“Society for Data Parasites”
Joe Gray

Gordon B Mills
The Jeff Drazen Controversy

DETAILS

However, many of us who have actually conducted clinical research, managed clinical studies and data collection and analysis, and curated data sets have concerns about the details. The first concern is that someone not involved in the generation and collection of the data may not understand the choices made in defining the parameters. Special problems arise if data are to be combined from independent studies and considered comparable. How heterogeneous were the study populations? Were the eligibility criteria the same? Can it be assumed that the differences in study populations, data collection and analysis, and treatments, both protocol-specified and unspecified, can be ignored?
DETAILS

• Is there sufficient data in the database/EHR to provide confidence in data

• Do we need clinical trial grade data including outcomes

• Who will pay for clinical trial grade follow-up
  – Must invest in obtaining clinical trial grade data

• Will clinical trial grade follow-up be sufficient outside of clinical trials
  – No randomization
  – Biases and vagaries of data
  – Failure of clinical trials to confirm
Molecular Epidemiology

- Retrospective analysis of clinical data
- Excellent curators
- Linked to large scale molecular assays
- Experience with big data
- Metanalysis

- TEN YEARS OF CASE CONTROL STUDY COULD NOT BE REPLICATED/REPRODUCED
- REPRODUCIBILITY IN SCIENCE
- UNKNOWN BIAS IN DATA
- DATA NOT DESIGNED TO “ANSWER” QUESTION ASKED
OVARIAN CANCER AS AN EXAMPLE

• For patients where CA125 is positive, CA125 is a surrogate for number of tumor cells
• Rate of decline in CA125 predicts patient outcome
• CA125 is routinely followed
• Outcome in ovarian cancer is defined by response to multiple rounds of therapy
• Ongoing study to determine ability of patient records used in our “Big data” program to evaluate clinical outcomes: blinded to CA125 levels
OVARIAN CANCER AS AN EXAMPLE

• Early conclusions
  – Imaging not performed at consistent intervals
    • Not linked to treatment protocol
    • Primarily performed at change in therapy
  – Physician and nursing notes do not correlate with CA125 or imaging data
    • Patient continues to do well
    • No change in tumor
    • Imaging (when available) and CA125 show marked increase in tumor size
  – No or limited correlation between “blinded” big data and patient response course by CA125
Challenges with health record for clinical outcomes

Patient continues to do well

Imaging 3 masses with largest 30 cm diameter

Treated elsewhere

Extensive disease

Patient continues to do well

No change in tumor

No clinical data available

No imaging

Surgery

Chemo Platinum Taxol

Chemo Doxil

Chemo

Platinum

Taxol

Doxil

Chemo

Platinum

Taxol

Deceased
I AM SAYING PORNOGRAPHY HURTS ANYONE WHO READS IT, GARBAGE IN, GARBAGE OUT.

Jerry Falwell
American Pastor
1933 - 2007
I AM SAYING

PORNOGRAPHY

HURTS ANYONE

WHO READS IT,

GARBAGE IN,

GARBAGE OUT.

Jerry Falwell
American Pastor
1933 - 2007
I AM SAYING
BIG DATA
HURTS ANYONE
WHO READS IT,
GARBAGE IN,
GARBAGE OUT.

Jerry Falwell
American Pastor
1933 - 2007